Diving Deep into the Motion Representation of Video-Text Models

Chinmaya Devaraj **Cornelia Fermüller** **Yiannis Aloimonos**
University of Maryland, College Park
(chinmayd,fermulcm,jyaloimo)@umd.edu

Abstract

Videos are more informative than images because they capture the dynamics of the scene. By representing motion in videos, we can capture dynamic activities. In this work, we introduce GPT-4 generated motion descriptions that capture fine-grained motion descriptions of activities and apply them to three action datasets. We evaluated several video-text models on the task of retrieval of motion descriptions. We found that they fall far behind human expert performance on two action datasets, raising the question of whether video-text models understand motion in videos. To address it, we introduce a method of improving motion understanding in video-text models by utilizing motion descriptions. This method proves to be effective on two action datasets for the motion description retrieval task. The results draw attention to the need for quality captions involving fine-grained motion information in existing datasets and demonstrate the effectiveness of the proposed pipeline in understanding fine-grained motion during video-text retrieval.

1 Introduction

Since the introduction of large-scale use of contrastive learning for image and text representation (Radford et al., 2021), various efforts have been made to build video-text models (Ni et al., 2022; Luo et al., 2022; Fang et al., 2021; Wang et al., 2021) to relate video to text. Videos provide a way to access the dynamics or motion in the scene that a single image cannot capture (Fermüller et al., 2018; Fermüller and Maynord, 2022; Dessalene et al., 2023). Motion in videos could be due to the action depicted, the effect of camera movement (for example, in egocentric action videos), or a combination of camera motion and action (Ogale et al.).

We investigate how existing video-text models perceive motion due to the action. For this work, we define motion in action videos as the movement of actors or the movement of actors and objects. The challenge is the lack of datasets that explicitly describe video motion. Figure 1 shows some examples of captions from ActivityNet (Krishna et al., 2017) and the MSR-VTT (Xu et al., 2016) dataset. Although verbs are included in captions of multimodal datasets like ActivityNet, MSR-VTT, Howto100M (Miech et al., 2019), Spoken Moments in Time (Monfort et al., 2021), a detailed description of motion is not available. This calls for the need to have an exclusive benchmark to evaluate how video-text models interpret and respond to motion descriptions.

We use the human action datasets Kinetics-400 (Kay et al., 2017), UCF-101 (Soomro et al., 2012), and HMDB-51 (Kuehne et al., 2011) to circumvent the lack of quality annotations of motion descriptions. The advantage of action datasets is that for every action label, we can obtain the corresponding characteristic motion of the action by using large language models like GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023), which, to a large degree, produce accurate descriptions of the all the actions in the datasets. Figure 2 shows some examples of the descriptions generated by GPT-4 and corresponding videos and original captions.

We evaluate several video-text models on the motion description retrieval task using the HMDB-51 and the UCF-101 datasets. We compare against human performance and show that all models fall far behind, raising questions about the design of video-text models and the role of quality captions in training better models to capture human motion. To address this question, we propose a method described in section 4 to investigate if providing better motion description captions helps video-text models understand fine-grained motion descriptions. To validate this fairly, we compare our method with video-text models that similarly initialize their video encoder and text encoder with pre-trained CLIP (Radford et al., 2021) weights so that undue

Findings of the Association for Computational Linguistics: ACL 2024, pages 12575–12584
August 11-16, 2024 ©2024 Association for Computational Linguistics

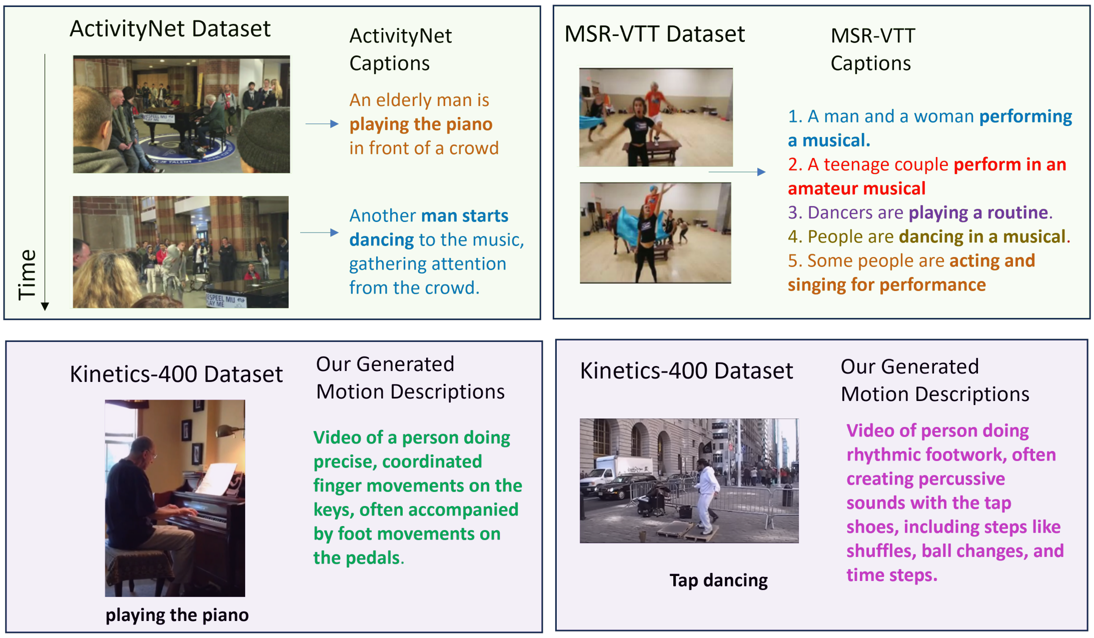

ActivityNet Dataset
ActivityNet Captions
Time
An elderly man is playing the piano in front of a crowd
Another man starts dancing to the music, gathering attention from the crowd.
MSR-VTT Dataset
MSR-VTT Captions
1. A man and a woman performing a musical.
2. A teenage couple perform in an amateur musical
3. Dancers are playing a routine.
4. People are dancing in a musical.
5. Some people are acting and singing for performance
Kinetics-400 Dataset
Our Generated Motion Descriptions
Video of a person doing precise, coordinated finger movements on the keys, often accompanied by foot movements on the pedals.
playing the piano
Kinetics-400 Dataset
Our Generated Motion Descriptions
Video of person doing rhythmic footwork, often creating percussive sounds with the tap shoes, including steps like shuffles, ball changes, and time steps.
Tap dancing

Figure 1: Example captions from ActivityNet, MSR-VTT, and our own GPT-4 generated fine-grained motion description for Kinetics-400 classes. Our generated motion descriptions solely describe the motion of the action, whereas other datasets typically use verbs to describe the scene.

performance gain is not obtained by pre-training on video-text data. Our results show that our proposed pipeline is very effective in learning fine-grained motion descriptions on both the UCF-101 dataset and the HMDB-51 dataset. In summary, our contributions are:

1. Creating a dataset of human motion descriptions for three action datasets.
2. Evaluating current video-text models representing motion description in videos on the UCF-101 and HMDB-51 datasets against human expert evaluation.
3. Introducing a method to validate the need for better captioning in video-text models to understand motion descriptions and demonstrate the method's effectiveness in capturing fine-grained motion descriptions.

2 Background and related work

Multimodal datasets and video understanding tasks: Howto100M and Spoken moments in time are popular video caption datasets used in pre-training video-text models. ActivityNet, MSR-VTT, DiDeMo (Hendricks et al., 2018), VaTex (Wang et al., 2019) are representative datasets used for video-language alignment tasks like video-to-text retrieval or text-to-video retrieval. To our knowledge, we are the first to introduce fine-grained motion descriptions in video datasets, which is not the focus of existing datasets. We introduce motion descriptions on Kinetics-400, HMDB-51, and UCF-101 datasets.

Video-text models: Various efforts have been made to build video-text models (Luo et al., 2022; Fang et al., 2021) mainly for video-to-text retrieval tasks. These models have developed mechanisms based on CLIP and extended them to video frames. Video-text models (Wang et al., 2021; Ni et al., 2022; Wu et al., 2023; Rasheed et al., 2023; Momeni et al., 2023a) have also been used for general video recognition both in the supervised and the zero-shot action recognition setting. (Momeni et al., 2023a) introduced verb-focused contrastive training to improve better verb reasoning by learning with hard negative verb examples. (Park et al., 2022) introduced contrast sets to identify pitfalls in video-text models and recommended the need for fine-grained action understanding to tackle hard negatives in contrast sets. We differ from them as we utilize motion descriptions, which has its unique challenge. We compare our video-text model with Vanilla CLIP (Radford et al., 2021), XCLIP (Ni et al., 2022), Text4Vis (Wu et al., 2023) and ViFiCLIP (Rasheed et al., 2023), primarily because they utilize the pre-trained CLIP for fair evaluation purposes.

3 Benchmark

Designing a motion description benchmark: We perform experiments on Kinetics-400, UCF-101, and HMDB-51 datasets. For each class in these action datasets, we prompt GPT-4 to produce the characteristic motions of the action (given by the caption annotation). Figure 2 shows the overall process of obtaining the motion descriptions. More details about the dataset statistics and generation are described in Appendix D. The dataset can be accessed at `https://github.com/chinmayad/motiondescriptions.git`

We conducted a user study to evaluate the quality of generated motion descriptions. Following (Karpinska et al., 2021), which questions the use of Amazon Mechanical Turk for such studies, we conducted this study with expert graduate student volunteers who have taken courses in computer vision and natural language processing. The evaluators were shown different questions in a training session. All the evaluators were trained with different questions and explained the project's overall goal and how they contributed to it. We asked two volunteer graduate students of different ages and ethnicities to participate in this study.

The following definitions were given.

1. Conciseness: Conciseness is related to the length and non-redundancy of the generated text.

2. Hallucinations: Hallucinations are related to generating physically non-plausible motion descriptions.

3. Relevance: Relevance refers to how much correspondence there is between the objects, action, and motion description.

4. Correctness: Correctness refers to how accurate the motion description is.

5. Harmfulness: Is there any objectionable or harmful content in the generated motion description?

Each of the above attributes is evaluated on a 5-point Likert scale. We report the mean 5-point Likert score and IAA%, the inter-annotator agreement that measures the percentage of descriptions where annotators gave the same rating. We asked the volunteers to rate the generated motion description with the above attributes for each motion description in each dataset. The study results are given in table1.

Method	Mean	IAA%
Conciseness	3.86	47.5
Hallucinations	1.12	19.35
Relevance	3.4	87
Correctness	3.92	72
Harmfulness	1	100

Table 1: Evaluation of the quality of generated motion descriptions

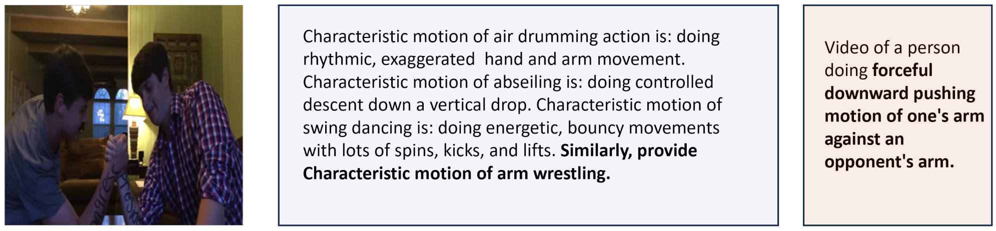
We noticed that many of the actions in the UCF-101 and HMDB-51 datasets involve objects, and the retrieval task is easier when an object is present in the generated motion description. For this reason, we also created another set of motion descriptions in which we replaced the names of objects with the generic word "object," which made the task slightly more challenging.

4 Proposed method

This section describes our proposed video-text model that incorporates the textual motion description. A typical vision-language model like CLIP is usually trained using contrastive learning loss or its variations (Miech et al., 2020; Momeni et al., 2023a). The quality of representations learned from the video-text model (Momeni et al., 2023b) will depend on the captions used during pre-training. Since no large video-caption dataset containing motion descriptions exists, we can't directly train a video-text model using contrastive learning. We, therefore, propose a method that utilizes the rich linguistical understanding of motion in actions from GPT-4 to give us the captions required for training. Our approach has two parts: Generating the required motion descriptions of actions using GPT-4 described in section 3 and training video-text models to utilize these motion descriptions described in section 4.2.

4.1 Problem setting

The model is trained on kinetics-400 as source dataset D_s and tested on target datasets D_t: UCF101 and HMDB51. The source dataset D_s consists of videos x_s and labels L_s belonging to classes C_s. The target dataset D_t consists of videos x_t and labels L_t belonging to classes C_t. We use a zero-shot setting such that $C_s \cap C_t = \emptyset$. Let the



Characteristic motion of air drumming action is: doing rhythmic, exaggerated hand and arm movement. Characteristic motion of abseiling is: doing controlled descent down a vertical drop. Characteristic motion of swing dancing is: doing energetic, bouncy movements with lots of spins, kicks, and lifts. **Similarly, provide Characteristic motion of arm wrestling.**

Video of a person doing **forceful downward pushing motion of one's arm against an opponent's arm.**

Video of **arm wrestling** from Kinetics-400 dataset.

Input Prompt to GPT-4

Generated motion description caption by GPT-4.

Figure 2: **Schematic representation of the generation of motion descriptions in existing action datasets.**

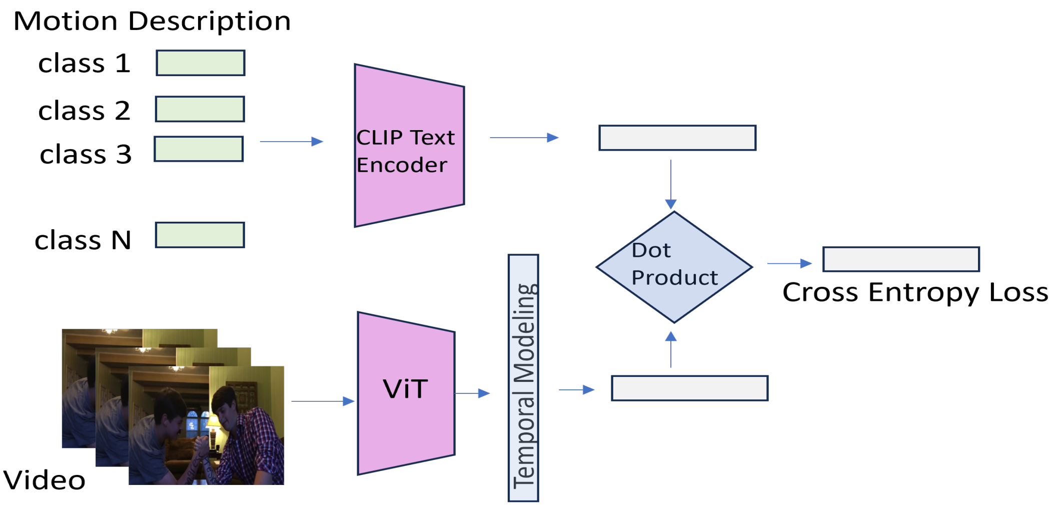


Figure 3: **Schematic representation of our approach encoding motion description in video-text model pipeline:** We integrate motion information as classifier weight in a supervised training paradigm. We finetune the image encoder to integrate the motion information while classifying videos in the kinetics400 dataset.

generated motion descriptions from GPT-4 for L_s and L_t be M_s and M_t respectively.

4.2 Architecture setting

Figure 3 gives an overview of our approach. While training, we freeze the text encoder from CLIP and fine-tune the visual encoder to learn the motion representation. While testing, motion descriptions are passed through the network to obtain classifier weights. Given a video from the target dataset, we obtain the logits indicating the probability that a video matches the corresponding motion description. A detailed discussion on training, testing, and implementation details is given in Appendix A.

4.3 Theoretical justification

We provide a theoretical justification for our proposed approach. Consider a large-scale dataset D containing visual samples with ground-truth labels. Denoting our labeled dataset $D = (x1, y1), (x2, y2), ...$, let X represent input data $x1, x2, ...$ and Y represent the labels $y1, y2, ...$

A typical supervised learning framework with a linear predictor involves minimizing $L(X^T W, Y) + Sigma(W)$ where W contains the parameters to be learned, $Sigma$ is the regularizing function, and L is the loss function. Here, W is learned independently on D and will not be helpful for new classes or other downstream datasets. As proposed in (Romera-Paredes and Torr, 2015), to make the approach tractable for zero-shot learning, we need to make W so it carries valuable information for new classes. The authors of (Romera-Paredes and Torr, 2015) introduce

$$W = VS^T, \qquad (1)$$

which we refer to as equation 1, where S is the signature of classes obtained from attributes of classes in D, and V is a new set of parameters to be learned.

For our scenario, we want to fine-tune the video-text model on the Kinetics dataset so that it can learn the motion description.

Let us denote V_E as the visual encoder from CLIP or any pre-trained video-text model. The supervised learning formulation to learn V_E^* and W_{proj}^* as in (Wu et al., 2023) can now be represented in minimizing cross-entropy $H(y|\sigma(W_{proj}.V_E(x)))$ referred to as equation 2, where $H(p*|p)$ stands for the Cross Entropy between the predicted distribution p and the ground-truth distribution $p*$. σ denotes the softmax operation, $W_{proj} \in R^{c \times d}$ denotes the linear projection matrix for classification where c is number of classes and d is the dimension of embedding from V_E. The above formulation in equation 2 is a standard visual feature transferring paradigm, where the visual encoder V_E. and the projection matrix (classifier) W_{proj} are learned simultaneously. We need to introduce a motion description to make the formulation in equation 2 learn the motion description and be useful for recognizing new motion descriptions for zero-shot settings.

Inspired by equation 1 where $W = VS^T$, we introduce motion description by making W_{proj} the signature of classes S, and V_E the new set of parameters of the visual encoder to be learned. In (Romera-Paredes and Torr, 2015) S was obtained from an attribute matrix, and in our work, we obtain W_{proj} as embeddings of a motion descriptor obtained from a CLIP text encoder.

Method	Object	Masked Object
Vanilla CLIP	25.92	23.33
Text4Vis	51.23	33.80
XCLIP	52.37	32.01
ViFiCLIP	52.70	34.76
Our Method	58.46	47.80
Human estimate	98	98

Table 2: Evaluation of percentage accuracy in motion description retrieval task on UCF-101 dataset.

5 Results

Task: For the target datasets UCF-101 and HMDB-51, the input is a video and the list of generated motion descriptions for all the classes in the dataset. The video-text model predicts the closest motion description that describes the video. The metric used is the percentage accuracy of correctly predicted motion descriptions.

The models we evaluate are Vanilla CLIP (Radford et al., 2021), XCLIP (Ni et al., 2022), Text4Vis (Wu et al., 2023) and VifiCLIP (Rasheed et al., 2023). Details about the models are given in Appendix B.

Method	Object	Masked Object
Vanilla CLIP	25.26	16.67
XCLIP	29.35	19.35
Text4Vis	34.12	24.93
VifiCLIP	36.20	28.9
Our Method	39.24	28.41
Human estimate	97.5	96

Table 3: Evaluation of percentage accuracy in motion description retrieval task on HMDB-51 dataset.

Human estimated performance: We sampled five videos randomly from each class in UCF-101 and HMDB-51. A human expert was asked to select the motion description that correctly describes the video from the list of descriptions generated by GPT4. Human experts are graduate students who have taken computer vision and NLP graduate courses and volunteered for this study.

Table 2 reports the performance of various approaches on the UCF101 dataset. Our proposed method beats previous methods by over 5% for motion descriptions containing the names of objects involved and by over 10% for motion descriptions where the word "object" replaces the object's name. We noticed that all the video-text models perform very poorly compared to human-estimated performance. We also see that video-text models have a strong bias toward nouns. When the specific name of the object involved is not used, there is an average 10% drop in performance for all methods, indicating the strong bias video-text models have for objects. Table 3 reports the performance of various approaches on the HMDB-51 dataset. Similar trends are found in the HMDB-51 dataset.

6 Conclusion

We introduced a benchmark to understand how motion is understood in video-text models. We highlighted the limitations in obtaining quality annotations describing motion in video. We also showed that the performance of video-text models for retrieving motion descriptions is poor compared to human expert performance. Our proposed method circumvents some of these issues and improves over other video-text models. While designing video-text models, we leave it to future work to build better models to capture motion and ignore the biases due to the object or the scene.

7 Limitations

We use a CLIP text encoder trained on image-text data to represent motion descriptions. This is not the best thing to do as, in practice, the CLIP text encoder would never have encountered the dynamics of videos while training. However, we hope the method works if we replace this CLIP text encoder with any other video-text model text encoder. Another limitation is that we trained on the Kinetics-400 dataset and tested on the UCF-101 and HMDB-51 datasets. As shown in the results section, the presence of an object or scene can often impact the performance during the retrieval task on these datasets. Furthermore, since we fine-tune image encoders on the source dataset, the possibility of overfitting the source dataset exists, leading to poor transferability on another target dataset. There are also potential biases in generated descriptions by GPT-4, and human quality estimation is expensive. For our experiments, volunteers spent a total of 16 hours.

8 Ethical Considerations

We aim to highlight the neglected aspect of modeling motion in video-text models. We think incorporating motion descriptions and reducing the biases of video-text models to objects and scenes positively impacts the design of video-text models. However, there could be a potential risk introduced in our method, as we rely on GPT-4 to provide us with motion descriptions of actions.

9 Acknowledgements

The support of NSF under awards OISE 2020624, BCS 2318255, and ARL under the Army Cooperative Agreement W911NF2120076 is greatly acknowledged.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. 2023. Therbligs in action: Video understanding through motion primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10626.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Cornelia Fermüller and Michael Maynord. 2022. Learning for action-based scene understanding. In *Advanced Methods and Deep Learning in Computer Vision*, pages 373–403. Elsevier.

Cornelia Fermüller, Fang Wang, Yezhou Yang, Konstantinos Zampogiannis, Yi Zhang, Francisco Barranco, and Michael Pfeiffer. 2018. Prediction of manipulation actions. *International Journal of Computer Vision*, 126:358–374.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023a. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15579–15591.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023b. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591.

Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. 2021. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063.

Abhitit Ogale, Cornelia Fermüller, and Yiannis Aloimonos. Motion segmentation using occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

OpenAI. 2023. Gpt-4 technical report.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Wenhao Wu, Zhun Sun, and Wanli Ouyang. 2023. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2847–2855.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

A Implementation details of our method

A.1 Problem Setting:

The model is trained on kinetics-400 as source dataset D_s and tested on target datasets D_t: UCF-101 and HMDB-51. The source dataset D_s consists of videos x_s and labels L_s belonging to classes C_s. The target dataset D_t consists of videos x_t and labels L_t belonging to classes C_t. We use a zero-shot setting such that $C_s \cap C_t = \emptyset$. Let the generated motion descriptions from GPT-4 for L_s and L_t be M_s and M_t respectively.

A.2 Training

Motion descriptions M_s are passed through a frozen CLIP text encoder to obtain the class prototypes of L_s. Our intuition is that these class prototypes can be approximated as classifier weights of the supervised video classifier. The concept takes its motivation from the work of (Nukrai et al., 2022; Liang et al., 2022), which shows that text embeddings from CLIP and vision embeddings from CLIP are very similar and fall within a ball of small radius. The obtained CLIP embeddings of motion descriptions from the frozen text encoder would approximately translate to visual class prototypes if obtained visually. With that intuition, we use the class prototypes from the frozen CLIP text encoder as the classifier weights of a visual classifier.

Given a video v_s from the source dataset D_s, T frames are sampled uniformly. The sampled frames are passed through a CLIP pre-trained Image encoder and temporally pooled to obtain a visual feature of the video. Then, the logits are obtained by computing the dot product of this video feature with the transpose of the classifier weights W_s. The model is trained using cross-entropy loss over logits and labels L_s, and the parameters of the CLIP image encoder are updated.

A.3 Testing

The motion descriptions M_t are passed through the network to obtain classifier weights W_t. Given a video v_t from the target dataset D_t, we obtain the logits indicating the probability that a video matches the corresponding motion description M_t.

A.4 Experimental details

Our model uses a VIT-B/16 pre-trained CLIP text and image encoder. We use eight frame samples per video. The CLIP text encoder was kept frozen during the training, and the CLIP image encoder was fine-tuned. We train the model for 10 epochs on the Kinetics-400 dataset with a learning rate of 0.00005 with a batch of the size of 20 on 4 NVIDIA RTX A5000 for 40 GPU hours. We use a learning warm step of 5 and a weight decay of 0.2. We use the Adam optimizer with the cross-entropy loss for training on the Kinetics dataset with a clip ratio of 0.1. Here, we describe more details about our baselines. We report the best results after running experiments on 5 runs.

A.5 Temporal modeling

We experimented with adding a 6-layer temporal transformer on the video head of the VIT-B/16 transformer. The results are shown below in table 4 and 5. Contrary to our initial hypothesis, having a temporal transformer didn't improve the performance over mean average pooling.

Method	Object	Masked Object
Temporal transformer	57.02	44.4
Mean Average Pooling	58.46	47.80

Table 4: Evaluation of percentage accuracy in motion description retrieval task on UCF-101 dataset.

Method	Object	Masked Object
Temporal transformer	37.53	26.84
Mean Average Pooling	39.24	28.41

Table 5: Evaluation of percentage accuracy in motion description retrieval task on HMDB-51 dataset.

A.6 Does fine-tuning cause overfitting?

As in any fine-tuning method, there is a risk of overfitting the source dataset. We performed experiments to see if overfitting is an issue. Based on our experiments, the degree to which the model overfits is negligible compared to the method's overall improvement. Table 6 and Table 7 below show the accuracies at different epochs of fine-tuning the vision encoder.

B Baselines

B.1 Vanilla CLIP:

We use VIT-B/16 pre-trained CLIP text and image encoder obtained from (Radford et al., 2021)

Number of Epochs	Object	Masked Object
Epoch 5	57.50	46.59
Epoch 10	58.46	47.80
Epoch 20	58.2	47.02

Table 6: Evaluation of percentage accuracy in motion description retrieval task on UCF-101 dataset.

Number of Epochs	Object	Masked Object
Epoch 5	38.25	27.82
Epoch 10	39.239	28.41

Table 7: Evaluation of percentage accuracy in motion description retrieval task on HMDB-51 dataset.

and temporally average the frame outputs while evaluating HMDB-51 and UCF-101 datasets.

B.2 XCLIP

We use the XCLIP (Ni et al., 2022) model and code available from https://huggingface.co/docs/transformers/model_doc/xclip. We use "microsoft/xclip-base-patch16-zero-shot" model from huggingface.co while evaluating the HMDB-51 and UCF-101 datasets.

B.3 Text4Vision

We use the VIT-B/16 base architecture with 8 frames per video. We use pre-trained weights from https://github.com/whwu95/Text4Vis/tree/main

B.4 VifiCLIP

We use the VIT-B/16 architecture and obtain the model and code from the official implementation of VifiCLIP (Rasheed et al., 2023) from https://github.com/muzairkhattak/ViFi-CLIP.

C Dataset

C.1 Kinetics-400

Kinetics-400 is a large-scale dataset containing 400 classes downloaded from YouTube. It has 240K training videos and 20K validation videos. Some videos are missing if the YouTube user has removed them.

C.2 UCF-101

The UCF-101 human action dataset consists of 13 K YouTube videos belonging to 101 classes. We report results on full classes on one split provided by the authors.

C.3 HMDB-51

It contains approximately 7K videos belonging to 51 classes. We report results on the split provided by the authors.

D Motion Description Generation

We use the GPT -4 API to obtain motion descriptions. The generated motion descriptions for the three datasets are provided as supplementary data along with this submission.

We noticed from experiments that we needed to provide some example motion descriptions in the prompt to obtain motion descriptions in the format we are interested in. The prompt we used is "*Characteristic motion of air drumming action is: doing rhythmic, exaggerated hand and arm movement. The characteristic motion of abseiling is: doing a controlled descent down a vertical drop. The characteristic motion of swing dancing is: doing energetic, bouncy movements with lots of spins, kicks, and lifts. Similarly, provide the characteristic motions of action X.*".

D.1 GPT-4 generated motion descriptions quality control

5 volunteers evaluated the generated motion descriptions for pairwise comparison. The pairwise comparison included selecting the best one among two generated motion descriptions. A vote of majority was used to select the final motion description. Volunteers with diverse experience and age groups were selected to reduce bias.

D.2 Dataset statistics

The kinetics-400 dataset consists of 400 classes, the UCF-101 human action dataset consists of 101 classes and HMDB-51 consists of 51 classes. We generate a characteristic motion description for each class in all three datasets. For UCF101 and HMDB-51, we mask objects manually after obtaining the motion description. Table D.2 provides the motion description dataset statistics.

Dataset	Number of videos	Number of unique motion descriptions	Average number of verbs per motion description	Average number of words per motion description	Number of verbs in the motion descriptions
Kinetics 400	246000	400	3.4	19	1371
UCF 101	13320	101	3.2	19	325
HMDB 51	6849	51	3.2	17	164

Table 8: Statistics of motion description dataset